Rapport sur la première campagne d'évaluation

Livrable D5

Projet PASSAGE

Référence: ANR-06-MDCA-013-02

Échéance : T12 - 31 décembre 2007)

Livraison: T18 - 30 juin 2008

Patrick Paroubek
LIMSI-CNRS
Bât. 508 Université Paris XI
91403 Orsay Cedex
tél. 01 69 85 80 04
email pap@limsi.fr

Contents

Chapitre 1

Le protocole d'évaluation

1.1 Contexte historique

La première campagne d'évaluation PASSAGE fait suite aux deux campagnes EASY[1] [13] du projet EVALDA[2] du programme TECHNOLANGUE. Ces premières campagnes ont permis de mettre en place un protocole d'évaluation prenant en compte de manière séquentielle :

- La création d'un large corpus de textes ;
- La définition d'un guide d'annotation pour les annotations syntaxiques;
- La constitution d'une annotation de référence sur une partie du corpus servant au test des systèmes ;
- L'utilisation du corpus de test par des systèmes participants ;
- L'évaluation des sorties de ces systèmes sur la partie annotée du corpus de test ;
- L'extension des annotations sur l'ensemble du corpus de test en fusionnant les sorties des analyseurs syntaxiques.

1.2 Le formalisme d'annotation

Les annotations syntaxiques comprennent d'un part des constituants (une suite de mots contigus motivée syntaxiquement) dont on distingue 6 sortes :

1. nominal,
2. adjectival,
3. prépositionnel,
4. adverbial,
5. verbal
6. et prépositionnel-verbal.

[1] Evaluation des Analyseurs Syntaxiques
[2] http://www.elda.org/rubrique69.html

Le dernier type de constituant est utilisé pour décrire les verbes à l'infinitif introduits par une préposition. Les constituants sont définis par leur portée (la suite de mot délimité par les adresses des premier et du dernier mots de la séquence) et une étiquette décrivant leur type.

D'autre part, les annotations comprennent des 14 sortes de relations syntaxiques de dépendance :

1. sujet-verbe,
2. auxiliaire-verbe,
3. complément d'objet direct,
4. complément-verbe,
5. modifieur de non,
6. modifieur de verbe,
7. modifieur d'adjectif,
8. modifieur d'adverbe,
9. modifieur de préposition,
10. complémenteur,
11. attribut du sujet/objet,
12. coordination,
13. apposition,
14. juxtaposition.

Le choix de ces constituants et de ces relations est le produit d'une discussion avec les participants au projet en partant du formalisme utilisé pour la campagne EASY du projet TECHNOLANGUE. Une description détaillée du formalisme est disponible dans le guide d'annotation [3], Ils sont également décrits dans [18]. La figure 1.1 donne un exemple d'annotation d'une phrase issue du corpus littéraire.

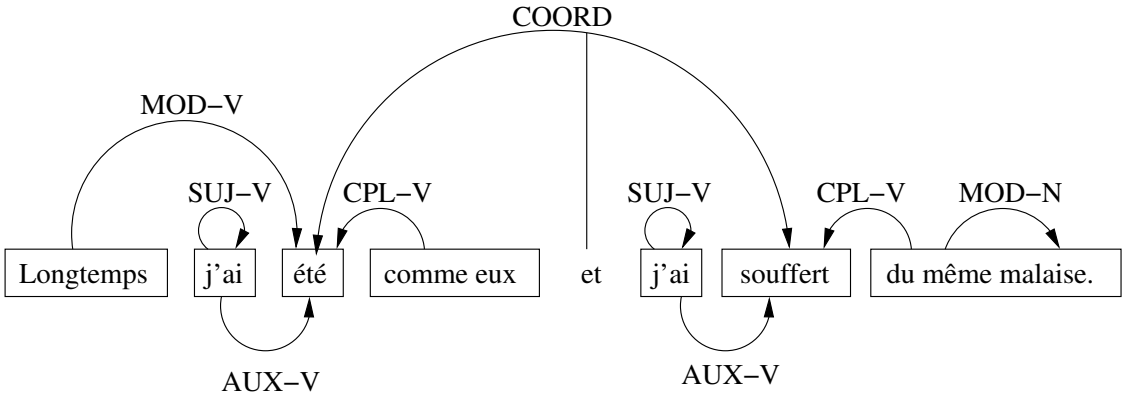

Figure 1.1: Exemple d'annotation d'un énoncé extrait du corpus littéraire (Coppé)

[3] Le guide d'annotation est disponible à l'URL www.limsi.fr/Recherche/CORVAL/easy

1.3 La procédure d'évaluation

L'évaluation des annotations peut se réaliser indépendamment en constituants (Groupe Nominal, Noyau Verbal, etc.) et/ou en relations (Sujet-Verbe, Attribut du sujet, etc.). De manière plus fine, les scores de précision, rappel et f-mesure [19] sont calculés spécifiquement par annotation (par exemple la relation sujet-verbe) et par genre textuel (journalistique, spécialité, email, transcription d'oral etc.), chaque genre correspondant à un sous-corpus spécifique.

Les mesures d'évaluation admettent 15 relâchements de contraintes différentes, obtenus en combinant les 5 manières de comparer les segments de textes correspondants aux constituants ou cibles de relations, avec les 3 façons de considérer les définitions des constituants (ceux présent dans les données hypothèse, ceux présent dans les données de référence, ou ceux de l'hypothèse s'ils existent sinon ceux de la référence).

Fonction	Formule
ÉGALITÉ	$H = R$
FLOU UNITAIRE	$\|H \backslash R\| \leq 1$
INCLUSION	$H \subset R$
INTERSECTION	$R \cap H \neq \emptyset$
BARYCENTRE	$\frac{2*\|R \cap H\|}{\|R\|+\|H\|} > 0.25$

Table 1.1: Avec H l'empan de texte hypothèse et R l'empan de texte référence, la table donne les formules permettant de comparer les empans correspondant soit aux constituants, soit aux sources/cibles de relations.

La première campagne d'évaluation PASSAGE comprend deux pistes d'évaluation, l'une appelée « EASY classique » et l'autre appelée PASSAGE. La première adhère strictement au protocole de la campagne d'évaluation EASY [12] et fixe pour les corpus une segmentation en mots et en phrases que les participants doivent obligatoirement respecter ; la seconde laisse libre la segmentation en mot et en phrase et repose sur un réalignement avec un algorithme de programmation dynamique [10] des données fournies par les participants pour calculer par vote majoritaire une segmentation en mots et en phrases commune. De plus, les annotations de référence pour la piste PASSAGE sont elles aussi obtenues en combinant les annotations des participants selon une stratégie de vote majoritaire ; la même qui sera utilisée pour annoter un corpus de plusieurs centaines de millions de mots, résultat de la campagne PASSAGE. Les deux pistes utilisent le format XML avec une DTD dérivée de celle de la campagne EASY.

Le corpus utilisé dans la piste « EASY classique » est le même que celui de la dernière campagne EASY. Le corpus textuel comprend 40 000 énoncés (généralement une phrase) dont un sous-ensemble de 4 000 énoncés annotées a servi de référence lors de la campagne EASY. Le corpus et ces annotations sont connus des participants et utilisés par eux pour améliorer les performances de leur système pendant la phase de développement. À ce corpus de 4 000 énoncés, un complément d'annotations manuelles pour 400 énoncés pris dans la partie du corpus EASY non encore annotée (partie du corpus EASY complémentaire de celle utilisée comme référence dans la campagne EASY) a été ajouté spécifiquement pour la piste « EASY classique » de la campagne PASSAGE. Ce complément de corpus permet de s'assurer, en partie, que les systèmes n'ont pas été « surentraîné » sur le corpus de développement en vérifiant à la fin de la phase de développement que les performances obtenues sur l'ancienne référence EASY et celles obtenues sur son complément apporté pour PASSAGE sont corrélées. Bien entendu, une meilleur validation de ce point aurait été obtenue si au lieu de se contenter de compléter les annotations de références en annotant des énoncés déjà connues des participants nous aurions apporté du nouveau matériau annoté au lieu de simplement apporter de nouvelles annotations ; dans une campagne d'évaluation le corpus de test doit toujours être disjoint du corpus de développement. Mais ce critère de validation nous a néanmoins paru suffire car d'une part, les participants n'ont de leur propre aveux pas eu le temps de modifier leur système depuis la fin de la campagne EASY et d'autre part, la suite des activités de PASSAGE après cette première campagne, n'utilisera plus de segmentations en mots et en phrases définies a priori. Pour ces raisons, il nous a paru opportun de nous contenter d'un test de validation sur des annotations complémentaires portant sur un matériau connu des participants.

C'est donc la première phase de développement qui est capitale et qui a dès lors nécessité la création d'une plateforme d'évaluation accessible depuis un serveur Web commun : en effet, un tel processus d'évaluation implique des expérimentations à répétition de la part des participants, devenant très rapidement fastidieuses à effectuer et nécessitant l'installation et l'utilisation de la chaîne d'évaluation.

Pour résumer, le corpus de référence de cette première campagne EASY se décompose en trois parties :

1. Le corpus annoté manuellement, utilisé en phase de développement (quantité moyenne, 4 000 énoncés dont le texte et les annotations sont connues des participants) ;

2. Le corpus complémentaire annoté manuellement, utilisé en phase de test (quantité faible, 400 énoncés dont seul le texte est connu des participants) ;

3. Le corpus non annoté (quantité importante, 40 000 énoncés), qui lui sera utilisé pour réaliser une annotation automatique de l'ensemble du corpus à l'aide de la fusion de l'ensemble des sorties de système selon un algorithme de vote majoritaire.

Dans la campagne PASSAGE, 13 systèmes d'analyse syntaxique ont été testés et tous ont apprécié les apports du serveur d'évaluation, à tel point que dès la phase de test terminée, ils ont spontanément demandé à ce que le serveur d'évaluation soit ré-ouvert car il souhaitaient poursuivre leurs expérimentations avec les corpus de référence et la chaîne d'évaluation.

Chapitre 2

Les données

2.1 Les systèmes des participants

Les 11 systèmes qui ont participé à la première campagne d'évaluation PASSAGE sont les suivant :

- FRMG, un analyseur hybride TIG/TAG dérivé d'une métagrammaire et développé à l'INRIA[1] [5], [15], [6];
- SXLFG, un analyseur LFG développé à l'INRIA [5], [4],
- LLP2 un analyseur TAG lui aussi dérivé d'une métagrammaire développé au LORIA[2] [14];
- LIMA un analyseur en dépendances développé au LIC2M / CEA-LIST[3] [2];
- TAGParser un chunker étendu développé par TAGMATICA[4] [8];
- Deux analyseurs basés sur les Grammaires de Propriétés, développés au LPL[5] qui fonctionnent par satisfaction de contraintes [3]. Le premier est un analyseur symbolique déterministe tandis que le second est un analyseur statistique entraîné sur les données de la campagne EASY [16]
- CORDIAL un analyseur à base de règles développé par SYNAPSE[6];
- SYGMART un analyseur developpé au LIRMM[7];
- XIP un analyseur à cascade de règles développé au Xerox Research Center Europe[8] [1].

2.2 Les corpus

Le détail des différents corpus utilisés lors de la première campagne PASSAGE est donné dans la table 2.2. La diversité des genre textuels présents dans le corpus permet de tester la solidité des analyseurs face aux variations qu'ils peuvent rencontrer dans les domaines associés aux genres textuels..

[1] http://www.inria.fr/rocquencourt
[2] http://www.loria.fr/
[3] http://www-list.cea.fr/
[4] http://www.tagmatica.com/
[5] http://cnrs.oxcs.fr/
[6] http://www.synapse-fr.com/
[7] http://www.lirmm.fr/xml/fr/lirmm.html
[8] http://www.xrce.xerox.com/

Ressource	mots	Description
WIKIPEDIA	200K	Un corpus à accès libre, constitué d'une œuvre collective, couvrant différents domaines de connaissance offrant des styles variés mais orientés vers les descriptions.
WIKINEWS	18.2K	Une collection de nouvelles journalistiques courtes.
WIKILIVRES	170K	Une collection à accès libre, de 1956 livres pédagogiques en français issus WIKIBOOKS.
EUROPARL	200K	Un corpus multilingue parallèle de textes extraits de la partie française des actes du parlement européen.
JRC-ACQUIS	120K	Une partie de l'ensemble des lois de l'Union Européenne, qui existe en plusieurs langues.
ESTER	100K	Un corpus de transcriptions orales provenant du projet ESTER [9]
LE MONDE	100K	Un corpus journalistique de nouvelles internationales.
EASY:	1M	Le corpus utilisé pour la campagne EASY qui contient différents genres littéraires et inclus un sous-ensemble d'environ 4K énoncés (76K mots) qui ont été validés manuellement ainsi que 400 énoncés dont les annotations ont été ajoutées à l'occasion de la première campagne PASSAGE.
- LE MONDE	86K	
- Parlementaire	82K	
- Littéraire	230K	Romans français
- Oral du DELIC	9K	transcription d'oral
- Oral d' Ester	12K	
- Médical	50K	textes médicaux de différents domaines
- Questions	52K	Questions de différentes sources
- Web	17K	pages web
- email	150K	

Table 2.1: Détail des corpus utilisés dans la campagne 1 de PASSAGE

2.3 Les résultats de la première campagne

2.3.1 Les mesures de constituants de la piste «classique» EASY

La figure 2.1 donne les résultats que les 10 systèmes qui ont participé à la tâche d'annotation des constituants ont obtenu.

Pour la plupart des systèmes la valeur de la F-mesure (combinaison des mesures de précision et de rappel) va jusqu'à 90% et seulement trois analyseurs ont obtenu une performance entre 80% et 90%. La tendance est la même pour les mesures de précision et de rappel. En considérant tous les types de constituants confondus, l'analyseur P05 est le meilleur système. Environ 96,5% des constituants qu'il retourne sont corrects et il a trouvé 95,5% des constituants présents dans la référence. La figure 2.2 donne les scores de performance en F-mesure obtenus par le système P05 pour chaque constituant et pour chaque genre de corpus.

Même si les scores de cet analyseur sont les meilleurs, il représente le comportement global des autres systèmes :

- les scores sont plus bas pour les constituants adjectivaux (GA) et adverbiaux (GR);
- les scores sont plus élevés pour les constituants verbaux, prépositionels et prépositionel-verbaux.

Néanmoins, notez que les performances de deux systèmes (P02 et P08) baisse de manière notoire pour les constituants prépositionel-verbaux (PV).

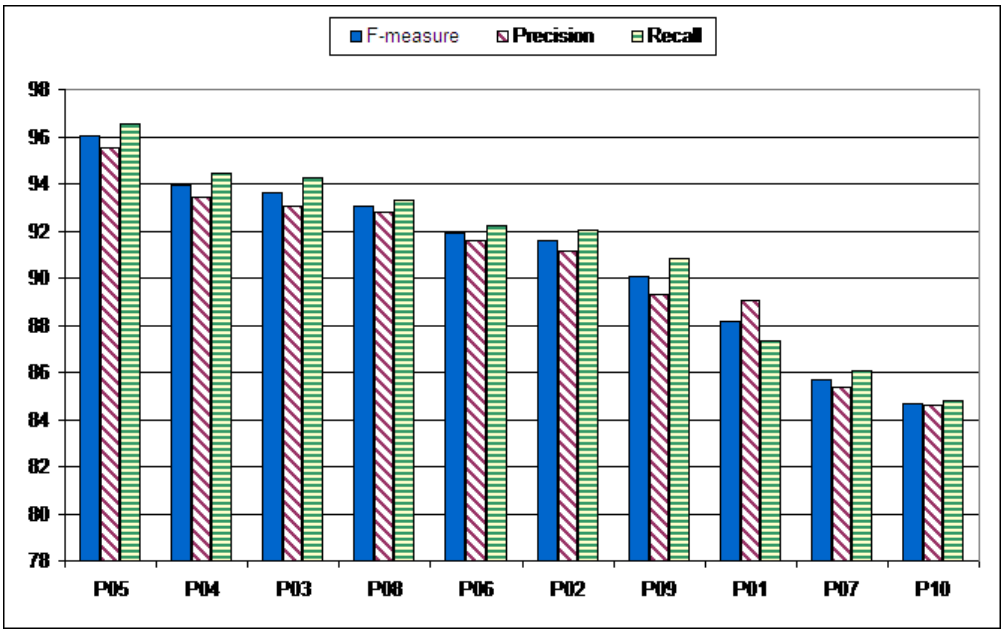

Figure 2.1: Performances en F-mesure en constituants pour tous les analyseurs (tous constituants confondus)

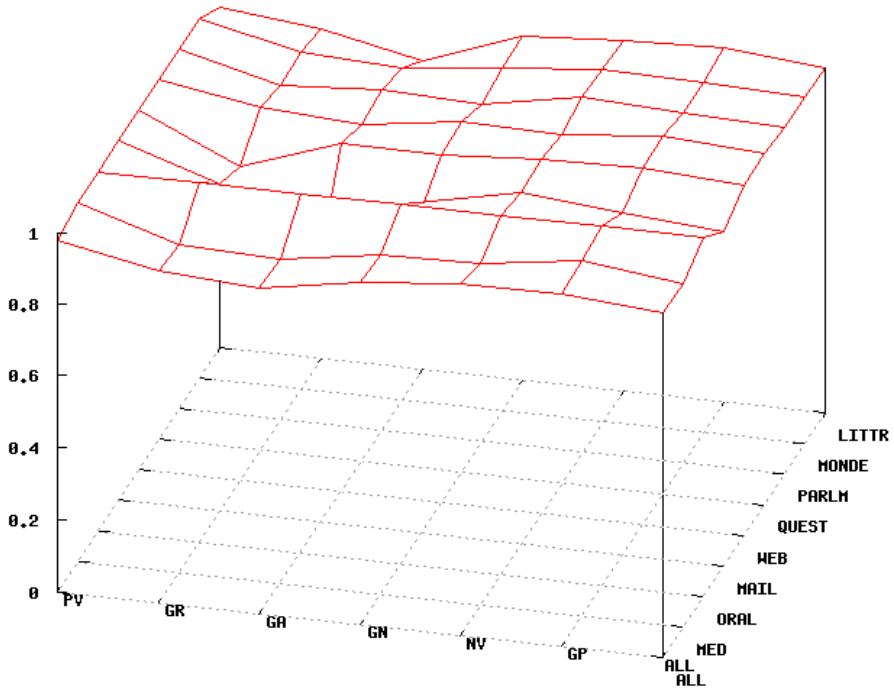

Figure 2.2: Performances en F-mesure par type de constituant et type de corpus pour l'analyseur P05

Les systèmes ont obtenus en général de bonnes performances avec les corpus *parlementaire* et *questions*, tandis que les corpus *le monde, littéraire* ou *médical* sont associés avec des résultats légèrement moins bons. Les scores pour le corpus *email* sont les plus bas. Il faut noter que les scores de chaque analyseur ne varient pas beaucoup en fonction du genre de corpus traité, par ex. l'analyseur P05 obtient 97.6% en F-mesure pour le corpus *questions* et 92.9% pour le corpus *email*. Ils ne varient pas non plus beaucoup en fonction du type de constituant, au final on obtient donc des courbes de performances très plates comme celle de la figure 2.2.

2.3.2 Les mesures sur les relations de la piste «classique» EASY

Les résultats que les 7 systèmes qui ont participé à la tâche d'annotation des relations ont obtenu sont donnés dans la figure 2.3.

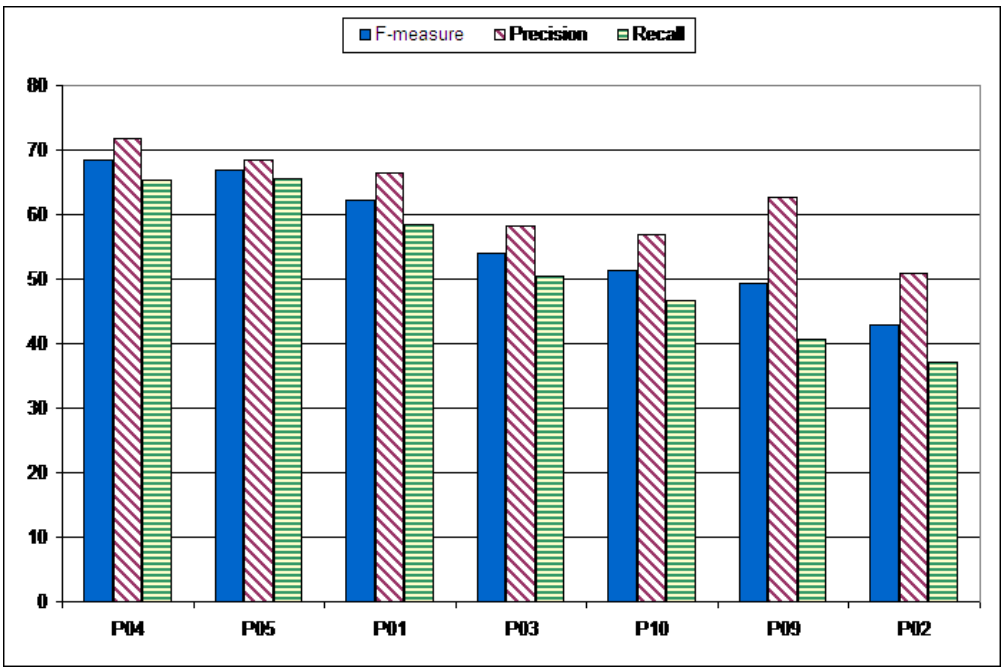

Figure 2.3: Performances en F-mesure concernant les relations pour tous les analyseurs (toutes relations confondues)

Bien sur les performances sont plus basses que pour les constituants et les différences de performance observées entre les analyseurs ont augmenté. Une conséquence évidente du fait que l'annotation en dépendances est une tâche beaucoup plus complexe que l'annotation en constituants. Aucun système n'a de performances supérieure à 70% en F-mesure, trois sont au dessus de 60% et deux au dessus de 50%. Les deux derniers systèmes sont au dessus de 40%. Le système P04 a obtenu les meilleurs résultats dont le détail est donné dans la figure 2.4 par type de relation et par genre de corpus.

Les performances changent beaucoup en fonction du type de relation, mais une tendance générale émerge :

- les systèmes ont peu de problèmes avec les relations auxiliaire/verbe (environ 96% en F-mesure pour P04), et pour une moindre part en ce qui concerne les relations de modifieur de nom (environ 77% de F-mesure pour P04) et sujet-verbe (environ 78%);
- pour certaines relations les scores sont très bas : par exemple la relation de modifieur d'adverbe (12% en F-mesure pour P04), modifieur de préposition (0%), l'apposition (9%), et la juxtaposition (5%);

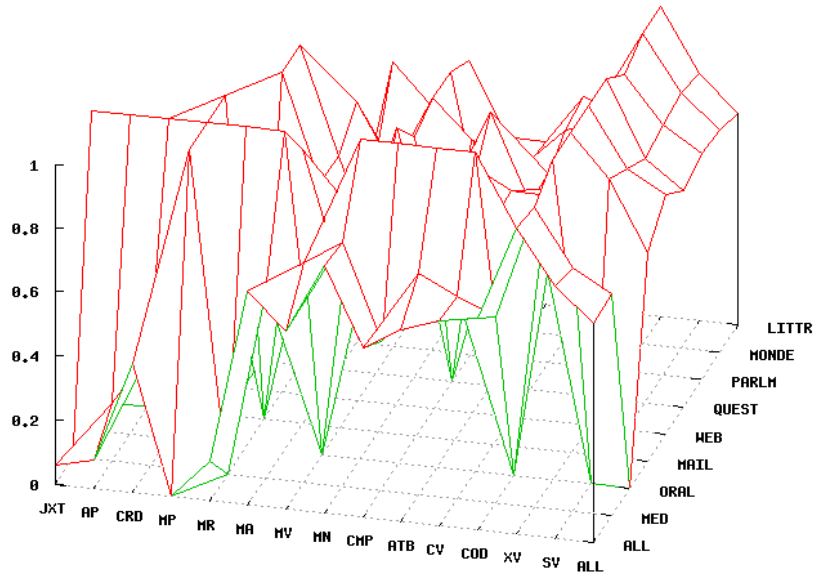

Figure 2.4: Performance en F-mesure pour les relations de l'analyseur P04 par type de relation et par genre de corpus.

- les pour les autres relations se situent entre 40% et 70% avec les systèmes P04, P05 et P01, mais sont inférieurs à 50% pour les autres.

Cette tendance se retrouve dans tous les corpus, par exemple l'analyseur P04 obtient des scores qui varient de 74.7% en F-mesure (pour le corpus *web*) à 64.9% (pour le corpus *email*).

2.3.3 Stabilité des performances en fonction du corpus

Pour observer la stabilité des résultats en fonction des différents corpus, nous avons calculé la variance de la F-mesure pondéré par la taille de la population. Les facteurs de pondération ont été extraits des informations présentées dans la table 2.2, qui donne le nombre d'énoncés, le nombre de constituants et le nombre de relations par corpus présents dans les données de référence. Nous avons ensuite comparé la variance pondérée à la valeur moyenne de la F-mesure.

Nous n'avons pas considéré le corpus oral_delic dans notre étude car il est trop petit (une conséquence des contraintes de temps qui ont limité l'annotation de données de référence pour cette première campagne). L'analyse de la stabilité des performances en fonction des corpus a été faite de la manière suivante. D'abord, nous avons calculé les variances pondérées par corpus pour chaque système d'une part pour les constituants et d'autre part pour les relations avec la formule suivante : $V = \frac{\sum_{i=1}^{N} (F_i - F_m)^2 * W_i}{N}$, ou $W_i = \frac{C_i}{C_{total}}$, N est le nombre total de corpus. F_i est la valeur de performance en F-mesure pour le corpus i pour les constituants (resp.. les relations), F_m est la valeur moyenne de F-mesure pour les constituants (resp. les relations), C_i est le nombre de constituants (resp. relations) présents dans les données de référence pour le corpus i et C_{total} est le nombre de constituants (resp. relations) dans les données de référence. Les deux types de score de variance sont donnés par système dans la figure Figure 2.5.

Les variances pondérées par corpus sont basses, ce qui signifie que les mesures de performance en F-mesures par système sont stables et qu'il n'y a pas de dispersion des scores au fil des différents corpus. Si nous excluons un système (P01, qui a une variance pour les constituants plus élevée que celle des autres

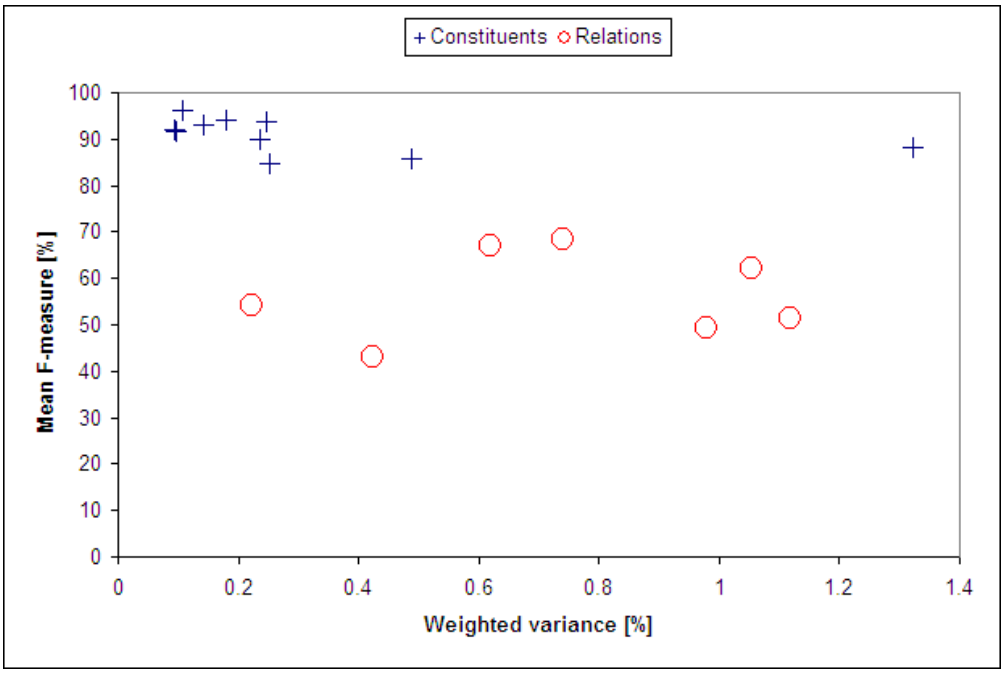

Figure 2.5: Variance pondéré par corpus pour chaque système

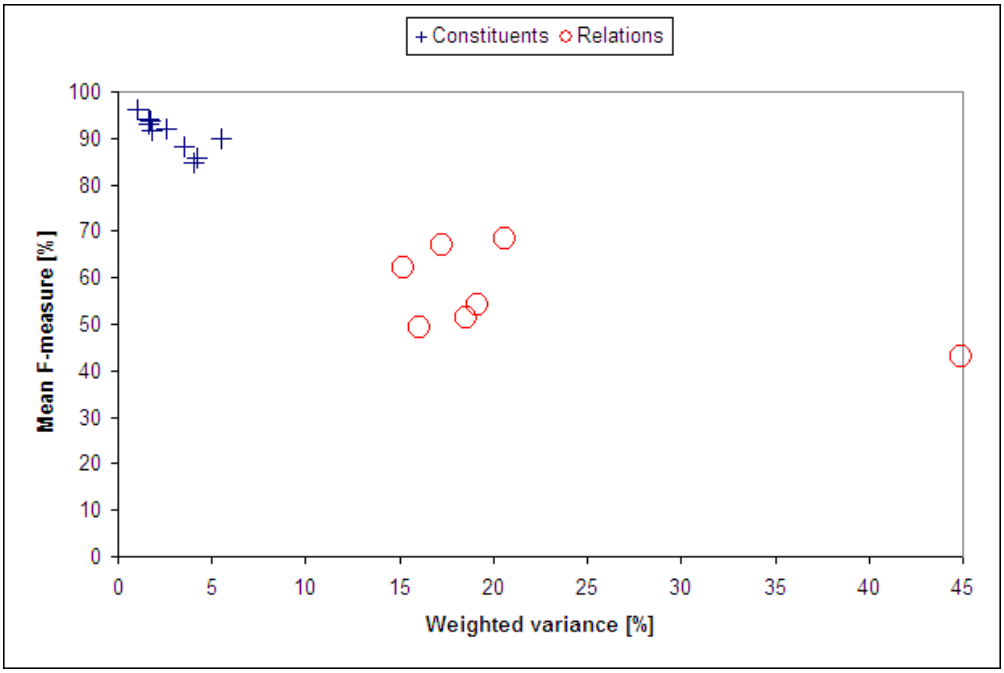

Figure 2.6: Variance pondérée par constituant/relation pour chaque système

systèmes), les variances pondérées par corpus sont plus basses pour les constituants que pour les relations.

Corpus	Énoncés	Constituants	Relations
lemonde	52	681	746
littéraire	163	1680	1971
email	21	194	240
médical	47	600	613
oral_delic	1	3	2
parlement	79	1093	1226
questions	28	252	257
web	14	177	183
Total	405	4680	5246

Table 2.2: Nombre d'énoncés, de constituants et de relation par corpus présents dans les données de référence de la première campagne PASSAGE.

De plus, les systèmes se comportent tous de la même façon si l'on regarde la variance pondérée pour les constituants. À la fois le corpus web (dans un sens positif) et le corpus email (dans un sens négatif) sont associés à des scores extrêmes, provoquant une dispersion des scores.

Lorsque l'on regarde les relations, les systèmes sont légèrement plus dispersés mais les systèmes se comportent sur les deux corpus précédents globalement de la même façon que précédemment. Remarquez que le système (P09) a obtenu des performances très basses sur le corpus de questions.

De manières similaire, nous avons calculé les variances par type de constituant ou de relation. Les variances sont alors pondérées par les nombre d'occurrence de chaque type des constituants au lieu du nombre global de constituants utilisés précédemment. Les résultats sont donnés dans la figure 2.6.

Les variances pondérées par type de constituant ou de relation sont beaucoup plus élevées que par type de corpus. Mais la aussi les systèmes montrent une tendance à l'aggrégation, à l'exception d'un système (P02 pour la variance en relations) et à se comporter de façon similaire. La dispersion des scores de F-measure sur les constituants est très élevée. tandis que la dispersion pour les relations est très élevée (cela correspond à un écart type pouvant aller jusqu'à 4).

Cette représentation permet de diagnostiquer plus en détails l'origine des problèmes que le systèmes ont rencontré :

- pour les constituants, la dispersion des scores des analyseurs provient des constituants adverbiaux (produisant des score bas) et des constituants prépositionels (induisant des scores élevés),
- pour les relations, la dispersion des scores des systèmes provient essentiellement d'une part des juxtapositions et des relations de complément du verbe pour les scores bas et d'autre part des relations subjet/verbe, auxilliaire/verbe et modifieur de nom pour les scores élevés.

Sur le graphique précédent nous pouvons noter que les relations comme la juxtaposition (dans un sens négatif) ou auxilliaire/verbe (dans un sens positif), sont trés excentrées et donc contribuent de manière très importantes à modifier la valeur moyenne de F-mesure.

2.4 Combinaison des analyses

La première étape de la combinaison des analyses sur le corpus de la track PASSAGE, celle qui consiste a combiner les segmentations en mots et en phrases selon une procédure de vote majoritaire, a produit la segmentation décrite dans la table 2.4. Dans le cas ou une portion de texte de taille trop importante n'est pas segmentée (pas assez d'analyse de segmentation disponibles ou divergence trop grapnde entre les données des analyseurs), l'algorithme recourt à une segmentation par défaut basée sur les espaces et les ponctuations fortes.

Corpus	Énoncés	Formes
ESTER	5.129	123.122
EUROPAR	8.406	223.898
WIKIPEDIA	11.194	173.853
WIKISOURCE	9.330	87.339
JRC	3.874	83.252
Le Monde	6.686	155.086
Total	44.619	846.550

Table 2.3: Table donnant le nombre d'énoncés et de formes par sous-corpus, obtenus par combinaison des segmentations en mots et phrases des analyseurs pour le corpus de la track PASSAGE.

La procédure de combinaison des annotations des analyseur n'est pas documentée ici, les travaux de mise au point des paramètres de combinaison n'étant pas pas encore achevés.

2.5 Publications

Les publications décrivant les résultats de la première campagne PASSAGE sont : [7] [11] et [17].

Bibliographie

[1] Ait-Mokhtar, S., Chanod, J.-P., Roux, C. 2002. Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering,* 8(3), 121–144.

[2] Besançon, R., de Chalendar, G. June 2005. L'analyseur syntaxique de lima dans la campagne d'évaluation easy. Dourdan, France. TALN'05.

[3] Blache, P. 2005. Property grammars: A fully constraint-based theory. Christiansen, H., (ed), *Constraint Solving and Language Processing* volume 3438. LNAI Springer.

[4] Boullier, P., Clément, L., Sagot, B., de la Clergerie, E. V. June 2005. Simple comme easy. Dourdan, France. TALN'05 57–60.

[5] Boullier, P., Sagot, B. 2005. Analyse syntaxique profonde à grande échelle: Sxlfg. *Traitement Automatique des Langues* 46(2), 65–89.

[6] de la Clergerie, E. V. October 2005. From metagrammars to factorized tag/tig parsers. Vancouver, Canada. IWPT'05.

[7] de la Clergerie, E. V., Hamon, O., Mostepha, D., Ayache, C., Paroubek, P., Vilnat, A. 2008. Passage: from french parser evalaution to large sized treebank. *Proceedings of the 6^{th} International Conference on Language Resources and Evaluation (LREC)* Marrakech, Morroco.

[8] Francopoulo, G. June 2005. Tagparser et technolangue-easy. Dourdan, France. TALN'05.

[9] Galliano, S., Geoffrois, E., Gravier, G., Bonastre, J.-F., Mostefa, D., Choukri, K. May 2006. Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. ELRA, , (ed), *In proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)* Genoa, Italy. ELRA.

[10] Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R. February 1999. Performance measures for information extraction. *Proceedings of DARPA Broadcast News Workshop* Herndnon VA.

[11] Olivier Hamon, D. M., Patrick Paroubek 2008. Sews un serveur d'évaluation orienté web pour la syntaxe. *Traitement Automatique du Langage* 49(2).

[12] Paroubek, P., Robba, I., Vilnat, A., Pouillot, L.-G. juin 2005. Easy : Campagne d'évaluation des analyseurs syntaxiques. *Actes des Ateliers de la $12^{ème}$ Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN 2005)* Dourdan.

[13] Paroubek, P., Vilnat, A., Robba, I., Ayache, C. juin 2007. Les résultats de la campagne easy d'évaluation des analyseurs syntaxiques du français. *Actes de la $14^{ème}$ Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN 2007)* Toulouse.

[14] Roussanaly, A., Crabbé, B., Perrin, J. June 2005. L'analyseur syntaxique de lima dans la campagne d'évaluation easy. Dourdan, France. TALN'05.

[15] Thomasset, F., de la Clergerie, E. V. June 2005. Comment obtenir plus des meta-grammaires. Dourdan, France. TALN'05.

[16] Vanrullen, T., Blache, P., Balfourier, J.-M. May 2006. Constraint-based parsing as an efficient solution: Results from the parsing evaluation campaign easy. ELRA, , (ed), *In proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)* Genoa, Italy. ELRA.

[17] Vilnat, A., Francopoulo, G., Hamon, O., Loiseau, S., Paroubek, P., de la Clergerie, E. V. August 2008. Large scale production of syntactic annotation to move forward. *Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation (PE 2008) in conjuction with COLING* Manchester. pp 36–43.

[18] Vilnat, A., Paroubek, P., Monceaux, L., Robba, I., Gendner, V., Illouz, G., Jardino, M. 2004. The ongoing evaluation campaign of syntactic parsing of french: Easy. *Proceedings of the 4^{th} International Conference on Language Resources and Evaluation (LREC)* Lisbonne, Portugal. 2023–2026.

[19] Vilnat, A., Paroubek, P., Monceaux, L., Robba, I., Gendner, V., Illouze, G., Jardino, M. June 2004. The ongoing evaluation campaign of syntactic parsing of french: Easy. *Proceedings of LREC* Lisboa, Portugal. 2023–2026.